ISSN:2319-2607 IISUniv.J.Sc.Tech. Vol.9(1), 24-28 (2020)
REVIEW ARTICLE

Study on Hadoop-MapReduce in Context of Cloud environment

Vandana Vijay*, Ruchi Nanda
Department of CS & IT, IIS (deemed to be University), Jaipur

Abstract
Cloud computing is changing the way people use computers and access, modify, or saves their personal as well as business information. It emerged as a new computational paradigm. It is an internet-based technology that enables any organization to use services offered over the Internet. Cloud computing has self-service provisioning, elasticity, and pay-per-use benefits. At the same time, it has many drawbacks (low scalability, security issues, no support for stream data processing). Hadoop paradigm has emerged as a universal tool for storing and processing huge data-sets. It handles virtually limitless concurrent tasks. It can process petabytes of data and run applications on thousands of nodes. Hadoop framework needs to be implemented in the cloud to overcome its drawbacks. This paper highlights the role of Hadoop in the context of the Cloud environment. Hadoop offers flexibility to the cloud to scale up or scale down as needs. It provides a flexible and agile computing platform to the cloud. It process workloads of structured and unstructured type on-demand. It handles batch workloads in the cloud efficiently. It can process terabytes of data in few minutes, and petabytes in hours. Hadoop enables businesses to run applications on thousands of nodes involving thousands of terabytes of data. It offers lower risk and worldwide availability.

Keywords: Big Data, Cloud computing, Hadoop, Hadoop Distributed File System, Hadoop Ecosystem, Indexing, MapReduce

Introduction

In this new period, everyone is moving their data, information, and applications to the cloud. Cloud provides high transfer speed. It has brought about the creation of gigantic information on an ordinary premise. The processing of this enormous information like the cloud is a troublesome assignment. For handling such a colossal measure of information, the conventional techniques, database management are not appropriate. Since these old methodologies neglected to deal with the tremendous size of information. Consequently, to deal with such large volume of information, organizations are offering multiple options. One of the accepted option is Hadoop, which possesses important features like flexibility, accessibility, scalable, and parallel processing. Hadoop works with the help of its two important components, MapReduce and Hadoop Distributed File System (HDFS). The paper is structured as: Section 1 covers Hadoop components (Distributed File System and MapReduce) and its ecosystem. Section 2 defines Cloud Computing and its service models (IaaS, PaaS, SaaS). Section 3 provides a literature review on research papers related to Cloud computing with Hadoop. Finally, Section 4 concludes this paper by highlighting the importance of Hadoop in the cloud computing environment and its future scope.

Hadoop

Hadoop is an Apache open-source framework. It can process an enormous volume of heterogeneous data. It provides distributed computing environment (Qayyum, 2020; Nagdive and Tugnayat 2018). Hadoop has two main components: HDFS and MapReduce. The storage of the large data within the distributed environment is supported by HDFS while the processing of the information in a distributed environment is processed by MapReduce (Vijay and Nanda 2020; Tripathi *et al.*, 2018; Jain and Alisha 2017). The MapReduce contains two tasks, namely Map and Reduce. (i) The Map task receives a set of data and change it into different set of data, where individual elements are broken down into tuples (key-value pairs). (ii) The Reduce task receives the output from the Map as an input and combines those data tuples (key-value pairs) into a smaller set of tuples as shown in Fig 1. Hadoop is capable of running MapReduce programs written in Java, Ruby, Python, and C++.

*Corresponding Author Email: vandanavijay161978@gmail.com

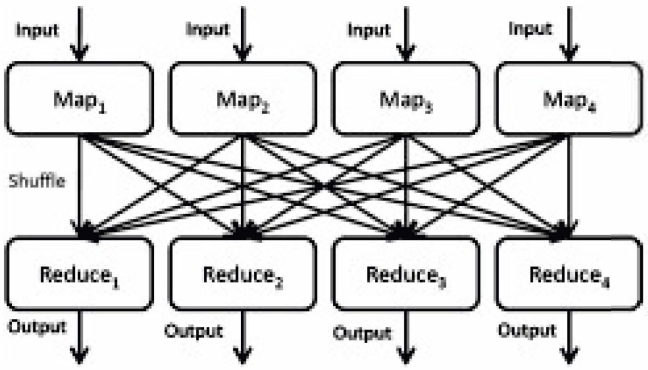

Fig. 1. MapReduce

Hadoop follows the master and slave concept (Qayyum, 2020). The NameNode, SecondaryNode, and Job Tracker belong to MasterNode, while the DataNode and Task Tracker belong to Slave. Job Tracker initiates the tasks and tracks them. Task Tracker process the input data and reports the states to Job Tracker. System metadata is managed by NameNode. DataNode stores actual data. It deals with reading/write requests from clients. Each DataNode also informs NameNode about the block it holds. These block reports are sent from time to time. If a DataNode fails, the error handling mechanism becomes active. HDFS is fault-tolerant. Hadoop ecosystem has a platform that solves large-scale computations. It includes commercial tools. Some of the popular tools include Spark, Yarn, Hive, Pig, Sqoop, Zookeeper, and Oozie which are listed as shown.

HDFS: Distributed File System

MapReduce: Distributed computation framework

YARN: Yet Another Resource Negotiator

Spark: In-memory Data Processing

Flume: Data Ingesting Services (data collection)

Zookeeper: Managing Cluster (Coordinator)

Apache Drill: SQL on Hadoop

PIG: Dataflow language

HIVE: Data warehouse infrastructure

Cloud Computing

Cloud computing is depicted as a model for giving on-request, network admittance to a shared pool of assets e.g., networks, servers, applications, and storage. Cloud is a network of servers that pools different resources (Nanda *et al.*, 2017). Cloud Providers offer services that can be categorized depending upon either the type of service being provided or based on location as shown in fig 2. The three basic service models are described (Ali *et al.*, 2021). 1. IaaS: Its full form is Infrastructure-as-a-service. Cloud suppliers provide computation resources, storage, and network as web-based services. 2. PaaS: Its full form is Platform-as-a-service. Cloud suppliers deliver platforms, tools, and other profits with the help of which one can create and deal with the applications, without the need of installing these platforms on the local machines. 3. SaaS: Its full form is Software-as-a-service. Cloud suppliers convey applications facilitated on the cloud foundation as web-based help for end-users, without requiring introducing the applications on the clients PCs.

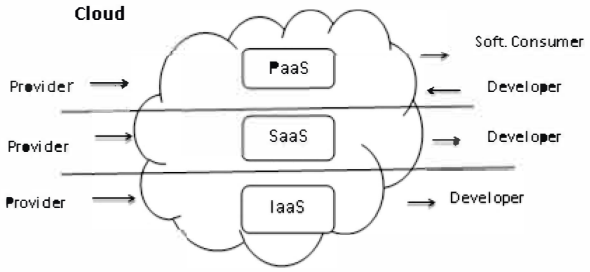

Fig. 2. Cloud computing

Importance of Hadoop Cluster in the Cloud

Hadoop clusters are run in the cloud due to the following reasons:

Scalability: Hadoop provides a highly scalable storage platform. It stores and operates large datasets across multiple servers which run in a parallel environment (Bashir, 2017; Li *et al.*, 2016). Hadoop can process petabytes of data.

Low Cost of Innovation: Running Hadoop on the cloud has the same preferred position as running some other programming offering on the cloud (Aher and Kulkarni 2015). The cloud additionally bodes well for a snappy, once use case including huge data processing.

Flexibility: There is no need to implement the servers physically. It is simpler to redesign examples or extend or contract the changing business requirements. As each task is controlled by web consoles and cloud provider APIs. (Alzakholi *et al.*, 2020; Maggiani, 2009).

Lack of space: When the user requires Hadoop clusters, but dont have space for servers. They can make use of the cloud (Gupta and Saxena 2014).

Lower risk: As per Patil *et al.*, 2016 users can easily get the number of usable resources. There is little risk of under commitment or over-commitment. If some resource malfunctions, that resource gets discarded and a new one is allocated.

Batch workloads are handled efficiently: Hadoop involves processing scheduled jobs for input data on a fixed, temporal basis in a batch-oriented system. Companies

collect data from devices or web servers and input this data into analytics applications on Hadoop. The cloud is more efficient to handle such batch workloads (Aher and Kulkarni 2015).

Distributed environment: Cloud is best for providing big data computation processing parallel (Qayyum, 2020; Tripathi *et al.*, 2018).

Hadoop cluster: As per Voruganti, 2014 Hadoop is designed to execute huge datasets by linking or binding multiple smaller nodes together. These nodes form a single, cost-effective cluster.

Review of Literature

This survey refers to the crux of the research papers published in the field of Hadoop concerning cloud computing. An attempt is made to throw light on the role of Hadoop in the Cloud environment and it has been presented here. Alzakholi et al.., 2020 compared cloud computing technologies and different methods to increase the performance of the cloud. Technologies include Hadoop, Dryad, CGL Map Reduce, High Energy Physics (HEP), and Cloud Burst. Bashir, 2017 reviews that cloud MapReduce has high scalability and simplifies large-scale data computation. Qayyum, 2020 presents Hadoop as a solution for processing and storage of semi-structured data types. Malhotra *et al.*, 2018 proposed the GENMR model. This model can process data at Cloud repositories so that the drawback of a traditional database system can be overcome. Tripathi *et al.*, 2018 developed a cloud-enabled Hadoop framework. It combines cloud technique and computing resources with the traditional Hadoop to support solutions related to big data. Ikhlaq and Keswani 2016 review Big Data methods, Approaches, and implementation of cloud computing. Patil *et al.*, 2016 proposed secured Hadoop as a cloud service. It resolves data security problems. They make use of an encryption/ decryption algorithm. They provide a way to process large data on the cloud with a security and hassle-free platform. Ansari *et al.*, 2015 proposed a Data Cleaning mechanism, Push Model, and caching in Hadoop. It reduces the computation time of big data. Voruganti, 2014 designed Hadoop to process big data by linking multiple nodes together. These nodes function as a parallel entity. Gupta and Saxena 2014 proposed Hadoop implementation in big data. They conclude that Hadoop is the most required technology for Cloud Computing. Cloud vendors offer more Hadoop clusters in business. Comparative analysis of the Review of Literature is shown in Table 1.

Conclusion

In this survey paper, the role of Hadoop-MapReduce in the context of the Cloud environment is investigated. Hadoop has been proven to be a useful tool for processing large data on multiple nodes on cloud platform. It ensures a robust and fault-tolerant system. Hadoop is the first choice for cloud computations due to the following reasons:

Open Source: Code can be modified or changed according to the requirements of the user.

Distributed Processing: Data is processed in parallel on a cluster of nodes.

Reliability: Data is reliably stored on the nodes regardless of machine failures.

Fault Tolerance: Failures of nodes or tasks are recovered automatically.

Scalability: New hardware can be easily appended in the already existing cluster.

Data Locality: Hadoop works on the data locality principle.

The researchers have worked upon the major features that enhance the performance of Hadoop such as Scalability, Low Cost of Innovation, Flexibility, Fault Tolerance. It is tabulated and shown in Table 2. Hadoop has become the most popular framework for processing big data in the cloud, however in the future, the following issues can be considered:

- A solution to handle heterogeneous data with security and better access on clouds is yet to be discovered.
- Future programming structures should permit customer frameworks to create powerful, versatile programming models that, while depending on parallel computation, abstract the details of parallelism.
- More, efficient ways can be found to minimize the time complexity of the system.

Table 1. Comparative analysis of Review of Literature

Authors	Developed/Proposed Methodology	Conclusion Drawn	Future Perspectives
Alzakholi *et al.*, 2020	Compared cloud computing technologies and different methods to increase the performance of the cloud. Technologies include Hadoop, Dryad, CGL Map Reduce, High Energy Physics (HEP), and Cloud Burst.	Hadoop is better and has the edge over all other technologies; the time of execution for Hadoop in most applications is faster because the number of the partition of data is significant. Hadoop is also more accessible for users, and it works with most operating systems.	When the data is enormous, all technologies have proximally the same features, but the Dryad is very slow than other technologies. More efforts are required in the future to enhance its performance.
Qayyum, 2020	Presents Hadoop as a best option for processing and storage of semi-structured data types.	MapReduce enables to scattered any type of data across the cluster of nodes through distributed and parallel processing.	In Future, work can be done on the requirement for installing Hadoop in a cloud server. It can integrate with new frameworks.
Tripathi *et al*, 2018	Cloud-enabled Hadoop framework proposed for processing big data in Distributed environment.	The utilization of Cloud Computing gives a better-dispersed framework to dealing with spatial huge information.	Future spatial applications, for example, cloud estimating, calamity appraisal, and smart city projects can be overseen and measure without any problem.
Bashir, 2017	MapReduce enhances the execution engine, for processing data easily. At the point when MapReduce is implemented on different nodes, it turns out to be quick and simple to get to access data on those nodes.	Cloud MapReduce implementation is faster. It provides high scalability. Large-scale data computation becomes easy. Multiple data can be processed in parallel nodes.	Optimization algorithms (genetic algorithm, ant colony optimization) can be used to optimize the MapReduce procedure in the future.
Ansari *et al.*, 2015	Proposed Data Cleaning mechanism, Push Model, and caching in Hadoop.	To increase the speed of the execution process, the already available data is cleaned by a Data cleaning mechanism.	In the future, more efficient ways can be found to minimize the time complexity of the system.
Zeebaree *et al.*, 2020	The performance of Hadoop in distributed systems is clearly very good compared with other software used for the same purpose.	Hadoop can hold petabytes of data and provides a large application in clusters. In addition, it store redundant data in different locations for guarantee data access.	
Voruganti, 2014	The proposed implementation of the MapReduce through two units: a JobTracker and many TaskTrackers.	Hadoop is designed to execute huge datasets by linking or binding multiple smaller nodes together.	Future programming structures should permit customer frameworks to create powerful, versatile models that can abstract the details of parallelism.
Malhotra *et al*, 2018	Proposed a model GENMR. It converts RDBMS queries to Map Reduce codes.	GENMR model can effectively process data at Cloud repositories.	The proposed model can be implemented in the future by using an existing programming language.
Gupta and Saxena 2014	Proposed big data implementation using Hadoop. Single node Hadoop cluster is set up.	Hadoop is the most required technology for Cloud Computing. As more Hadoop clusters are offered by cloud vendors in many businesses.	New Hadoop advancements should be simpler for users to work and to get information in and out. Subsequently, this incorporates direct access with standard conventions utilizing existing devices and strategies.
Ikhlaq and Keswani 2016	Review Big Data methods, Approaches, and cloud computing implementation. It also describes the challenges posed by it.	Cloud Computing "is a ray of h ope and can achieve great heights when applied on BigData to unveil the wealth of knowledge.	A solution to handle heterogeneous data with security and better access on clouds is yet to be discovered.
Patil *et al.*, 2016	Proposed secured Hadoop as a cloud service. It offers optimized utilization. It also provides opportunistic provisioning of cycles from idle nodes to different processes.	Secured Hadoop as a service on the infrastructure clouds will process big data on the cloud. It offers data security and a hassle-free platform.	

Table 2. Major features used in different studies for performance enhancement of Hadoop

S.No	Authors	Scalability	Low Cost of Innovation	Flexibility	Fault Tolerance	Distributed environment
1.	Alzakholi *et al.*, 2020			√		
2.	Tripathi *et al.*, 2018					√
3.	Jain and Alisha, 2017					√
4.	Malhotra *et al.*, 2018				√	
5.	Zeebaree *et al.*, 2020				√	
6.	Bashir, 2017	√				
7.	Li *et al.*, 2016	√				
8.	Aher and Kulkarni, 2015		√			
9.	Patil *et al.*, 2016				√	
10.	Maggiani, 2009			√		
11.	Gupta and Saxena, 2014		√			
12.	Qayyum, 2020					√
13.	Voruganti, 2014					√

References

Aher, S. B., Kulkarni, A. R. (2015) Hadoop MapReduce: A programming model for large scale data processing. *Am J Comp Sci Eng Surv* 3: 01-10.

Ali, M. H., Hosain, M. S., Hossain, M. A. (2021) Big Data Analysis using BigQuery on Cloud Computing Platform. *Australian J of Eng Inno Tech* **3**(1): 1-9.

Alzakholi, O., Shukur, H., Zebari, R., Abas, S., Sadeeq, M. (2020) Comparison among cloud technologies and cloud performance. *J of Appli Sci-Tech Trends* **1**(2): 40-47.

Ansari, S. M., Chepuri, S., Wadhai, V. (2015) Efficient Map Reduce Model with Hadoop Framework for Data Processing. *J Comp Sci Mob Comp* **4**: 691-696.

Bashir, B. (2017) An Approach of MapReduce Programming Model for Cloud Computing. *Int J Adv Res Comp Sci* **8**(2):43-45.

Gupta, N., Saxena, K. (2014) Cloud computing techniques for big data and Hadoop implementation. *Int J Eng Res Tech* **3**(4): 722-726.

Ikhlaq, S., Keswani, B. (2016) Computation of Big Data in Hadoop and Cloud Environment *IOSR J Eng* **6**(1): 31-39.

Jain, E. P., Alisha, E. (2017) A brief review on Hadoop architecture and its issues. *Inter J Eng Res Gen Sci* **5**(2): 211-217.

Li, Z., Yang, C., Liu, K., Hu, F., Jin, B. (2016) Automatic scaling Hadoop in the cloud for efficient process of big geospatial data. *ISPRS Inter J Geo-Infor* **5**(10): 173 https://doi.org/10.3390/ijgi5100173

Malhotra, S., Doja, M. N., Alam, B., Alam, M. (2018) Generalized Query Processing Mechanism in Cloud Database Management System. In Big Data Analytics, Springer, Singapore, 641-648. https://doi.org/10.1007/978-981-10-6620-7_61

Nagdive, A. S., Tugnayat, R. M. (2018) A Review of Hadoop ecosystem for big data. *Inter J Comp Appl* **180**(14): 35-40.

Patil, A. U., Patil, R. U., Pande, A. P., Patil, B. S. (2016) Secured Hadoop as A Service Based on Infrastructure Cloud Computing Environment. *Inter J Adv Res Comp Comm Eng* **3**(5): 1086-1090.

Qayyum, R. (2020) A Roadmap Towards Big Data Opportunities, Emerging Issues and Hadoop as a Solution. *Inter J Edu Mngt Eng (IJEME)* **10**(4): 8-17.

Maggiani, R (2009) Cloud Computing is Changing How We Communicate *IEEE Inter Prof Comm Conf, IPCC*, 1-4. https://DOI: 10.1109/PCC.2009.5208703

Nanda, R., Sharma, K.S., Chande, S. (2017) Determining Appropriate Cache-size for Cost-effective Cloud Database Queries. *Inter J Comp Apple* **157**: 29-34.

Tripathi, A. K., Agrawal, S., Gupta, R. D. (2018) A Comparative Analysis of Conventional Hadoop with Proposed Cloud-Enabled Hadoop Framework for Spatial Big Data Processing. *ISPRS Anna Photo Rem Sens Spat Infor Sci* **45**: 425-430.

Vijay, V., Nanda, R. (2020) Query Caching Technique Over Cloud-Based MapReduce System: A Survey. *In Risi Thre Exp Apple Solu, Springer, Singapore* **1187**: 19-25.

Voruganti, S. (2014) Map Reduce a Programming Model for Cloud Computing Based On Hadoop Ecosystem. *Inter J Comp Sci Infor Tech* **5**(3): 3794-3799.

Zeebaree, S. R., Shukur, H. M., Haji, L. M., Zebari, R. R., Jacksi, K., Abas, S. M. (2020) Characteristics and analysis of Hadoop distributed systems. *Techn Rep Kans Uni* **62**(4): 1555-1564.